

# Matching Dependencies with Arbitrary Attribute Values: Semantics, Query Answering and Integrity Constraints\*

Jaffer Gardezi  
University of Ottawa, SITE  
Ottawa, Canada  
jgard082@uottawa.ca

Leopoldo Bertossi<sup>†</sup>  
Carleton University, SCS  
Ottawa, Canada  
bertossi@scs.carleton.ca

Iljuju Kiringa  
University of Ottawa, SITE  
Ottawa, Canada  
kiringa@site.uottawa.ca

## ABSTRACT

Matching dependencies (MDs) were introduced to specify the identification or matching of certain attribute values in pairs of database tuples when some similarity conditions are satisfied. Their enforcement can be seen as a natural generalization of entity resolution. In what we call the *pure case* of MDs, any value from the underlying data domain can be used for the value in common that does the matching. We investigate the semantics and properties of data cleaning through the enforcement of matching dependencies for the pure case. We characterize the intended clean instances and also the *clean answers* to queries as those that are invariant under the cleaning process. The complexity of computing clean instances and clean answers to queries is investigated. Tractable and intractable cases depending on the MDs and queries are identified. Finally, we establish connections with database *repairs* under integrity constraints.

## 1. INTRODUCTION

A database instance may contain several tuples and values in them that refer to the same external entity that is being modeled through the database. In consequence, the database may be modeling the same entity in different forms, as different entities, which most likely is not the intended representation. This problem could be caused by errors in data, by data coming from different sources that use different formats or semantics, etc. In this case, the database is considered to contain dirty data, and it must undergo a cleansing process that goes through two interlinked phases: detecting tuples (or values therein) that should be matched or identified, and, of course, doing the actual matching. This problem is usually called *entity resolution*, *data fusion*, *du-*

*plicate record detection*, etc. Cf. [13, 10] for some recent surveys and [6] for recent work in this area.

Quite recently, and generalizing entity resolution, [14, 15] introduced *matching dependencies* (MDs), which are declarative specifications of matchings of attribute values that should hold under certain conditions. MDs help identify duplicate data and enforce their merging by exploiting semantic knowledge expressed.

Loosely speaking, an MD is a rule defined on a database which states that, for any pair of tuples from given relations within the database, if the values of certain attributes of the tuples are similar, then the values of another set of attributes should be considered to represent the same object. In consequence, they should take the same values. Here, similarity of values can mean equality or a domain-dependent similarity relationship, e.g. related to some metric, such as the edit distance.

*Example 1.* Consider the following database instance of a relation  $P$ .

Name	Phone	Address
John Smith	723-9583	10-43 Oak St.
J. Smith	(750) 723-9583	43 Oak St. Ap. 10

Similarity of the names in the two tuples (as measured by, e.g. edit distance) is insufficient to establish that the tuples refer to the same person. This is because the last name is a common one, and only the first initial of one of the names is given. However, similarity of their phone and address values indicates that the two tuples may be duplicates. This is expressed by an MD which states that, if two tuples from  $P$  have similar address and phone, then the names should match. In the notation of MDs, this is expressed as

$$P[\text{Phone}] \approx P[\text{Phone}] \wedge P[\text{Address}] \approx P[\text{Address}] \rightarrow P[\text{Name}] \Rightarrow P[\text{Name}]. \quad \square$$

The identification in [14, 15] of a new class of dependencies and their declarative formulation have become important additions to data cleaning research. In this work we investigate matching dependencies, starting from and refining the model-theoretic and dynamic semantics of MDs introduced in [15].

Any method of querying a dirty data source must address the issue of duplicate detection in order to obtain accurate answers. Typically, this is done by first cleaning the data by discarding or combining duplicate tuples and standardizing formats. The result will be a new database where the entity conflicts have been resolved. However, the entity resolution problem may have different *solution instances* (which we will

\*Research supported by the NSERC Strategic Network on Business Intelligence (BIN,ADC05) and NSERC/IBM CRDPJ/371084-2008.

<sup>†</sup>Faculty Fellow of the IBM CAS. Also affiliated to University of Concepción (Chile).

simply call *solutions*), i.e. different clean versions of the original database. The model-theoretic semantics that we propose and investigate defines and characterizes the class of solutions, i.e. of intended clean instances.

After a clean instance has been obtained, it can be queried as usual. However, the query answers will then depend on the particular solution at hand. So, it becomes relevant to characterize those query answers that are invariant under the different (sensible) ways of cleaning the data, i.e. that persist across the solutions. This is an interesting problem *per se*. However, it becomes crucial if one wants to obtain semantically clean answers while still querying the original dirty data source.

This kind of virtual cleaning and query answering on top of it have been investigated in the area of *consistent query answering* (CQA) [3], where, instead of MDs, classical integrity constraints (ICs) are considered, and database instances are *repaired* in order to restore consistency (cf. [9, 7, 11] for surveys of CQA). Virtual approaches to robust query answering under entity resolution and enforcement of matching dependencies are certainly unavoidable in virtual data integration systems.

In this paper we make the following contributions, among others:

1. We revisit the semantics of MDs introduced in [15], pointing out sensible and justified modifications of it. A new semantics for MD satisfaction is then proposed and formally developed.
2. Using the new MD semantics, we formally define the intended solutions for a given, initial instance,  $D_0$ , that may not satisfy a given set of MDs. They are called *minimally resolved instances* (MRIs) and are obtained through an iteration process that stepwise enforces the satisfaction of MDs until a stable instance is reached. The resulting instances minimally differ from  $D_0$  in terms of number of changes of attribute values.
- This semantics (and the whole paper) considers the *pure case* introduced in [15], in the sense that the values than can be chosen to match attribute values are arbitrarily taken from the underlying data domains. No matching functions are considered, like in [6], for example (where entire tuples are merged, not individual attribute values).
3. We introduce the notion of *resolved answers* to a query posed to  $D_0$ . They are the answers that are invariant under the MRIs.
4. We investigate the computability and complexity of computing MRIs and resolved answers, identifying syntactic conditions on MDs and conjunctive queries under which the latter becomes tractable via query rewriting. The rewritten queries are allowed to contain counting and transitive closure (recursion).
5. We identify cases where computing (actually, deciding) resolved answers is coNP-complete.

6. We establish a connection between MRIs and database repairs under functional dependencies as found in CQA. In the latter case, the repairs consider, as usual, a notion of minimality based on deletion of whole tuples and comparison under set inclusion. This reduction

allows us to profit from results in CQA, obtaining additional (in)tractability results for MDs.

This paper is organized as follows. Section 2 presents basic concepts and notations needed in the rest of the paper. Section 3 identifies some problems with the MD semantics, and refines it to address them. It also introduces the resolved instances and resolved answers to a query. Section 4 considers the problems of computing resolved instances and resolved query answers. Section 5 identifies queries and sets of MDs for which computing resolved answers becomes tractable via query rewriting. Section 6 establishes the connection with CQA. Section 7 presents some final conclusions.

## 2. PRELIMINARIES

In general terms, we consider a relational schema  $\mathcal{S}$  that includes an enumerable infinite domain  $U$ . An instance  $D$  of  $\mathcal{S}$  can be seen as a finite set of ground atoms of the form  $R(\bar{t})$ , where  $R$  is a database predicate in  $\mathcal{S}$ , and  $\bar{t}$  is a tuple of constants from  $U$ . We assume that each database tuple has an identifier, e.g. an extra attribute that acts as a key for the relation and is not subject to updates. In the following it will not be listed, unless necessary, as one of the attributes of a database predicate. It plays an auxiliary role only, to keep track of updates on the other attributes.  $R(D)$  denotes the extension of  $R$  in  $D$ . We sometimes refer to attribute  $A$  of  $R$  by  $R[A]$ . If the  $i$ th attribute of predicate  $R$  is  $A$ , for a tuple  $t = (c_1, \dots, c_j) \in R(D)$ ,  $t[A]$  denotes the value  $c_i$ . The symbol  $t[\bar{A}]$  denotes the vector whose entries are the values of the attributes in the vector  $\bar{A}$ . The attributes may have subdomains that are contained in  $U$ . Constants will be denoted by lower case letters at the beginning of the alphabet.

A matching dependency [14], involving predicates  $R(A_1, \dots, A_n)$ ,  $S(B_1, \dots, B_m)$ , is a rule of the form

$$\bigwedge_{i \in I, j \in J} R[A_i] \approx_{ij} S[B_j] \rightarrow \bigwedge_{i \in I', j \in J'} R[A_i] = S[B_j]. \quad (1)$$

Here  $R$  and  $S$  could be the same predicate.  $I, I'$  and  $J, J'$  are fixed subsets of  $\{1, \dots, n\}$  and  $\{1, \dots, m\}$ , resp. We assume that, when  $A_i, B_j$  are related via  $\approx_{ij}$  or  $=$  in (1), they share the same (sub)domain, so their values can be compared by the domain-dependent binary similarity predicate,  $\approx_{ij}$  or can be identified, resp.

The similarity operators, generically denoted with  $\approx$ , are assumed to have the properties of: (a) Symmetry: If  $x \approx y$ , then  $y \approx x$ . (b) Equality subsumption: If  $x = y$ , then  $x \approx y$ .

The MD in (1) is implicitly universally quantified in front and applied to pairs of tuples  $t_1, t_2$  for  $R$  and  $S$ , resp. The expression  $\bigwedge R[A_i] \approx_{ij} S[B_j]$  states that the values of the attributes  $A_i$  in tuple  $t_1$  are similar to those of attributes  $B_j$  in tuple  $t_2$ . If this holds, the expression  $R[A_i] = S[B_j]$  indicates that, for the same tuples  $t_1$  and  $t_2$ ,  $t_1[A_i]$  and  $t_2[B_j]$  on the RHS should be updated so that they become the same, i.e. their values are identified or matched. However, the attribute values to be used for this matching are left unspecified by (1).

For abbreviation, we will sometimes write MDs as

$$R[\bar{A}] \approx S[\bar{B}] \rightarrow R[\bar{C}] = S[\bar{E}], \quad (2)$$

where  $\bar{A}, \bar{B}, \bar{C}$ , and  $\bar{E}$  represent the lists of attributes,  $(A_1, \dots, A_k)$ ,  $(B_1, \dots, B_k)$ ,  $(C_1, \dots, C_{k'})$ , and  $(E_1, \dots, E_{k'})$ , respectively. We refer to the pairs of attributes  $(A_i, B_i)$  and

$(C_i, E_i)$  as *corresponding pairs* of attributes of the pairs  $(\bar{A}, \bar{B})$  and  $(\bar{C}, \bar{E})$ , respectively. For an instance  $D$  and a pair of tuples  $t_1 \in R(D)$  and  $t_2 \in S(D)$ ,  $t_1[\bar{A}] \approx t_2[\bar{B}]$  indicates that the similarities of the values for all corresponding pairs of attributes of  $(\bar{A}, \bar{B})$  hold. Similarly,  $t_1[\bar{C}] = t_2[\bar{E}]$  denotes the equality of the values of all pairs of corresponding attributes of  $(\bar{C}, \bar{E})$ .

Since an MD involves an update operation, the MD is a condition that is satisfied by a pair of database instances: an instance  $D$  and its updated instance  $D'$ .

**Definition 1.** [15] Let  $D, D'$  be instances of schema  $\mathcal{S}$  with predicates  $R$  and  $S$ , such that, for each tuple  $t$  in  $D$ , there is a unique tuple  $t'$  in  $D'$  with the same identifier as  $t$ , and viceversa. The pair  $(D, D')$  satisfies the MD  $m$  in (2), denoted  $(D, D') \models_F m$ , iff, for every pair of tuples  $t_R \in R(D)$  and  $t_S \in S(D)$ , if  $t_R$  and  $t_S$  satisfy  $t_R[\bar{A}] \approx t_S[\bar{B}]$ , then for the corresponding tuples  $t'_R$  and  $t'_S$  in  $R(D'), S(D')$ , resp., it holds: (a)  $t'_R[\bar{C}] = t'_S[\bar{E}]$ , and (b)  $t'_R[\bar{A}] \approx t'_S[\bar{B}]$ .  $\square$

Intuitively,  $D'$  in Definition 1 is an instance obtained from  $D$  by enforcing  $m$  on instance  $D$ . For a set  $M$  of MDs, and a pair of instances  $(D, D')$ ,  $(D, D') \models_F M$  means that  $(D, D') \models_F m$ , for every  $m \in M$ .

An instance  $D'$  is *stable* [15] for a set  $M$  of MDs if  $(D', D') \models_F M$ . Stable instances correspond to the intuitive notion of a clean database, in the sense that all the expected value identifications already take place in it. Although not explicitly developed in [15], for an instance  $D$ , if  $(D, D') \models_F M$  for a stable instance  $D'$ , then  $D'$  is expected to be reached as a fix-point of an iteration of value identification updates that starts from  $D$  and is based on  $M$ .

### 3. MD SEMANTICS REVISITED

Condition (b) in Definition 1 is used to avoid that the identification updates destroy the original similarities. Unfortunately, enforcing the requirement sometimes leads to counterintuitive results.

**Example 2.** Consider the following instance  $D$  with string-valued attributes, and MDs:

$R$	$A$	$B$	$C$
	$a$	$c$	$g$
	$a$	$c$	$ksp$

$S$	$E$	$F$
	$h$	$c$
	$msp$	$c$

$$R[A] \approx R[A] \rightarrow R[C] = R[C] \quad (3)$$

$$R[C] \approx S[E] \rightarrow R[B] = S[F] \quad (4)$$

For two strings  $s_1$  and  $s_2$ ,  $s_1 \approx s_2$  if the edit distance  $d$  between  $s_1$  and  $s_2$  satisfies  $d \leq 1$ . To produce an instance  $D'$  satisfying  $(D, D') \models_F M$ , the strings  $g$  and  $ksp$  must be changed to some common string  $s'$ .

Because of the similarities  $h \approx g$  and  $ksp \approx msp$ ,  $s'$  must be similar to the  $E$  attribute values of the tuples in  $S$ , by condition (b) of Definition 1 and MD (4). Clearly, there is no  $s'$  that is similar to both  $h$  and  $msp$ . Therefore, at least one of  $h$  and  $msp$  must be modified to some new value in  $D'$ .  $\square$

Another problem with the semantics of MDs is that it allows duplicate resolution in instances that are already resolved. Intuitively, there is no reason to change the values in an instance that is stable for a set of MDs  $M$ , because there is no

reason to believe, on the basis of  $M$ , that these values are in error. However, even if an instance  $D$  satisfies  $(D, D) \models_F M$ , it is always possible, by choosing different common values, to produce a different instance  $D'$  such that  $(D, D') \models_F M$ . This is illustrated in the next example.

**Example 3.** Let  $D$  be the instance below and the MD  $R[A] \approx R[A] \rightarrow R[B] = R[B]$ .

$R$	$A$	$B$
	$a$	$c$
	$a$	$c$

Although  $D$  is stable,  $(D, D') \models_F m$  is true for any  $D'$  where the  $B$  attribute values of the two tuples are the same.  $\square$

### 3.1 MD satisfaction

We now propose a new semantics for MD satisfaction that disallows unjustified attribute modifications. We keep condition (a) of Definition 1, while replacing condition (b) with a restriction on the possible updates that can be made.

**Definition 2.** Let  $D$  be an instance of schema  $\mathcal{S}$ ,  $R \in \mathcal{S}$ ,  $t_R \in R(D)$ ,  $C$  an attribute of  $R$ , and  $M$  a set of MDs. Value  $t_R[C]$  is *modifiable* if there exist  $S \in \mathcal{S}$ ,  $t_S \in S(D)$ , an  $m \in M$  of the form  $R[\bar{A}] \approx S[\bar{B}] \rightarrow R[\bar{C}] = S[\bar{E}]$ , and a corresponding pair  $(C, E)$  of  $(\bar{C}, \bar{E})$ , such that one of the following holds: 1.  $t_R[\bar{A}] \approx t_S[\bar{B}]$ , but  $t_R[C] \neq t_S[E]$ . 2.  $t_R[\bar{A}] \approx t_S[\bar{B}]$  and  $t_S[E]$  is modifiable.  $\square$

**Example 4.** Consider two relations  $R$  and  $S$  with two MDs defined on them:

$R$	$A$	$B$
$t_0$	$a_0$	$b$
$t_1$	$a_1$	$b$
$t_2$	$a_2$	$b$

$S$	$C$	$E$
$t_3$	$a_3$	$c$
$t_4$	$a_4$	$c$
$t_5$	$a_5$	$c$

$$m_1 : R[A] \approx R[A] \rightarrow R[B] = R[B],$$

$$m_2 : R[A] \approx S[C] \rightarrow R[B] = S[E].$$

The following similarities hold on the distinct constants of  $R$  and  $S$ :  $a_i \approx a_{(i+1) \bmod 6}$ ,  $0 \leq i \leq 5$ . The values  $t_2[B]$  and  $t_3[E]$  are modifiable by condition 1 of Definition 2,  $m_2$ ,  $a_2 \approx a_3$ , and  $t_2[B] \neq t_3[E]$ . For the same reason,  $t_0[B]$  and  $t_5[E]$  are modifiable.

Value  $t_1[B]$  is modifiable by condition 2 of Definition 2,  $m_1$ ,  $a_1 \approx a_2$ , and the fact that  $t_2[B]$  is modifiable. Similarly,  $t_4[E]$  is modifiable.  $\square$

**Definition 3.** Let  $D, D'$  be instances for  $\mathcal{S}$  with the same tuple ids,  $M$  a set of MDs, and  $m \in M$ .  $(D, D')$  satisfies  $m$ , denoted  $(D, D') \models m$ , iff:

- For any pair of tuples  $t_R \in R(D)$ ,  $t_S \in S(D)$ , if there exists an MD in  $M$  of the form  $R[\bar{A}] \approx S[\bar{B}] \rightarrow R[\bar{C}] = S[\bar{E}]$  and  $t_R[\bar{A}] \approx t_S[\bar{B}]$ , then for the corresponding tuples  $t'_R \in R(D')$  and  $t'_S \in S(D')$ , it holds  $t'_R[\bar{C}] = t'_S[\bar{E}]$ .
- For any tuple  $t_R \in R(D)$  and any attribute  $G$  of  $R$ , if  $t_R[G]$  is not modifiable, then  $t_R[G] = t_R[G]$ .  $\square$

Notice that the notion of satisfaction of an MD is relative to a set of MDs to which the former belongs (due to the modifiability condition). Of course, for a single MD  $m$ , we can consider the set  $M = \{m\}$ . Condition 2. captures a

natural default condition of persistence of values: those that have to be changed are changed only.

The definition of satisfaction of a set  $M$  of MDs,  $(D, D') \models M$ , is as usual. Also, as before, we define *stable* instance for  $M$  to mean  $(D, D) \models M$ . Except where otherwise noted, these are the notions of satisfaction and stability that we will use in the rest of this paper.

*Example 5.* Consider again example 4. The set of all  $D'$  such that  $(D, D') \models M$  is the set of all instances obtained from  $D$  by changing all values of  $R[B]$  and  $S[E]$  to a common value, and leaving all other values unchanged. This is because the values of  $R[B]$  and  $S[E]$  are the only modifiable values, and these values must be equal by condition 1 of Definition 3 and the given similarities.  $\square$

Condition 2 in Definition 3 on the set of updatable values does not prevent us from obtaining instances  $D'$  that enforce the MD, as the following theorem establishes.

*Theorem 1.* For any instance  $D$  and set of MDs  $M$ , there exists a  $D'$  such that  $(D, D') \models M$ . Moreover, for any attribute value that is changed from  $D$  to  $D'$ , the new value can be chosen arbitrarily, as long as it is consistent with  $(D, D') \models M$ .  $\square$

The new semantics introduced in Definition 3 solves the problems mentioned at the beginning of this section. Notice that it does not require additional changes to preserve similarities (if the original ones were broken). Furthermore, modifications of instances, unless required by the enforcement of matchings as specified by the MDs, are not allowed. Also notice that the instance  $D'$  in Theorem 1 is not guaranteed to be stable. We address this issue in the next section.

Moreover, as can be seen from the proof of Theorem 1, the new restriction imposed by Definition 3 is as strong as possible in the following sense: Any definition of MD satisfaction that includes condition 1. must allow the modification of the modifiable attributes (according to Definition 2). Otherwise, it is not possible to ensure, for arbitrary  $D$ , the existence of an instance  $D'$  with  $(D, D') \models M$ .

### 3.2 Resolved instances

According to the MD semantics in [15], although not explicitly stated there, a clean version  $D'$  of an instance  $D$  is an instance  $D'$  satisfying the conditions  $(D, D') \models M$  and  $(D', D') \models M$ . Due to the natural restrictions on updates captured by the new semantics (cf. Definition 3), the existence of such a  $D'$  is not guaranteed. Essentially, this is because  $D'$  is the result of a series of updates. The MDs are applied to the original instance  $D$  to produce a new instance, which may have new pairs of similar values, forcing another application of the MDs, which in their turn produces another instance, and so on, until a stable instance  $D'$  is reached. The pair  $(D, D')$  may not satisfy  $M$ . However, we will be interested in those instances  $D'$  just mentioned. The idea is to relax the condition  $(D, D') \models M$ , and obtain a stable  $D'$  after an iterative process of MD enforcement, which at each step, say  $k$ , makes sure that  $(D_{k-1}, D_k) \models M$ .

*Definition 4.* Let  $D$  be a database instance and  $M$  a set of MDs. A *resolved instance* for  $D$  wrt  $M$  is an instance  $D'$ , such that there is a finite (possibly empty) sequence of instances  $D_1, D_2, \dots, D_n$  with:  $(D, D_1) \models M$ ,  $(D_1, D_2) \models M, \dots, (D_{n-1}, D_n) \models M$ ,  $(D_n, D') \models M$ , and  $(D', D') \models M$ .  $\square$

Note that, by Definition 3, for an instance  $D$  satisfying  $(D, D) \models M$ , it holds  $(D, D') \models M$  if and only if  $D' = D$ . In this case, the only possible set of intermediate instances is the empty set and  $D$  is the only resolved instance. Thus, a resolved instance cannot be obtained by making changes to an instance that is already resolved.

*Theorem 2.* Given an instance  $D$  and a set  $M$  of MDs, there always exists a resolved instance of  $D$  with respect to  $M$ .  $\square$

*Example 6.* Consider the following instance  $D$  of a relation  $R$  and set  $M$  of MDs:

$R(D)$	$A$	$B$	$C$
	$a$	$b$	$d$
	$a$	$c$	$e$
	$a$	$b$	$e$

$$R[A] \approx R[A] \rightarrow R[B] \rightleftharpoons R[B],$$

$$R[B] \approx R[B] \rightarrow R[C] \rightleftharpoons R[C].$$

All pairs of distinct constants in  $R$  are dissimilar. Two resolved instances  $D_1$  and  $D_2$  of  $R$  are shown.

$R(D_1)$	$A$	$B$	$C$
	$a$	$b$	$d$
	$a$	$b$	$d$
	$a$	$b$	$d$

$R(D_2)$	$A$	$B$	$C$
	$a$	$b$	$e$
	$a$	$b$	$e$
	$a$	$b$	$e$

Notice that  $(D, D_1) \not\models M$ , because the value of the  $C$  attribute of the second tuple is not modifiable in  $D$ .  $\square$

The notion of resolved instance is one step towards the characterization of the intended clean instances. However, it still leaves room for refinement. Actually, the resolved instances that are of most interest for us are those that are somehow closest to the original instance. This consideration leads to the concept of *minimal resolved instance*, which uses as a measure of change the number of values that were modified to obtain the clean database. In Example 6, instance  $D_2$  is a minimal resolved instance, whereas  $D_1$  is not.

*Definition 5.* Let  $D$  be an instance.

(a)  $T_D := \{(t, A) \mid t \text{ is the id of a tuple in } D \text{ and } A \text{ is an attribute of the tuple}\}$ .

(b)  $f_D : T_D \rightarrow U$  is given by:  $f_D(t, A) :=$  the value for  $A$  in the tuple in  $D$  with id  $t$ .

(c) For an instance  $D'$  with the same tuple ids as  $D$ :

$$S_{D,D'} := \{(t, A) \in T_D \mid f_D(t, A) \neq f_{D'}(t, A)\}. \quad \square$$

Intuitively,  $S_{D,D'}$  is the set of all values changed in going from  $D$  to  $D'$ .

*Definition 6.* Let  $D$  be an instance and  $M$  a set of MDs. A *minimally resolved instance* (MRI) of  $D$  wrt  $M$  is a resolved instance  $D'$  such that  $|S_{D,D'}|$  is minimum, i.e. there is no resolved instance  $D''$  with  $|S_{D,D''}| < |S_{D,D'}|$ . We denote by  $Res(D, M)$  the set of minimal resolved instances of  $D$  wrt the set  $M$  of MDs.  $\square$

*Example 7.* Consider the instance below and the MD  $R[A] \approx S[C] \rightarrow R[B] \rightleftharpoons S[D]$ .

$R$	$A$	$B$
	$a_1$	$b_1$

$S$	$C$	$D$
	$c_1$	$d_1$



Assuming that  $a_1 \approx c_1$ , this instance has two minimal resolved instances, namely

$R$	$A$	$B$
	$a_1$	$d_1$

$S$	$C$	$D$
	$c_1$	$d_1$

$R$	$A$	$B$
	$a_1$	$b_1$

$S$	$C$	$D$
	$c_1$	$b_1$

□

Considering that MDs concentrate on changes of attribute values, we consider that this notion of minimality is appropriate. The comparisons have to be made at the attribute value level. Notice that in CQA a few other notions of minimality and comparison of instances have been investigated [7].

### 3.3 Resolved answers

Let  $\mathcal{Q}(\bar{x})$  be a query expressed in the first-order language  $L(S)$  associated to schema  $S$ . Now we are in position to characterize the admissible answers to  $\mathcal{Q}$  from  $D$ , as those that are invariant under the matching resolution process.

**Definition 7.** A tuple of constants  $\bar{a}$  is a *resolved answer* to  $\mathcal{Q}(\bar{x})$  wrt the set  $M$  of MDs, denoted  $D \models_M \mathcal{Q}[\bar{a}]$ , iff  $D' \models \mathcal{Q}[\bar{a}]$ , for every  $D' \in \text{Res}(D, M)$ . We denote with  $\text{ResAn}(D, \mathcal{Q}, M)$  the set of resolved answers to  $\mathcal{Q}$  from  $D$  wrt  $M$ . □

**Example 8.** (example 7 continued) The set of resolved answers to the query  $\mathcal{Q}_1(x, y) : R(x, y)$  is empty since there are no tuples that are in the instance of  $R$  in all minimal resolved instances. On the other hand, the set of resolved answers to the query  $\mathcal{Q}_2(x) : \exists y(R(x, y) \wedge (y = b_1 \vee y = d_1))$  is  $\{a_1\}$ . □

In Section 4 we will study the complexity of the problem of computing the resolved answers, which we now formally introduce.

**Definition 8.** Given a schema  $S$ , a query  $\mathcal{Q}(\bar{x}) \in L(S)$ , and a set  $M$  of MDs, the *Resolved Answer Problem* (RAP) is the problem of deciding membership of the set

$$RA_{\mathcal{Q}, M} := \{(D, \bar{a}) \mid \bar{a} \text{ is a resolved answer to } \mathcal{Q} \text{ from instance } D \text{ wrt } M\}.$$

If  $\mathcal{Q}$  is a boolean query, it is the problem of determining whether  $\mathcal{Q}$  is true in all minimal resolved instances of  $D$ . □

## 4. COMPUTING RESOLVED INSTANCES AND ANSWERS

In this section, we consider the complexity of the  $RA_{\mathcal{Q}, M}$  problem introduced in the previous section. For this goal it is useful to associate a graph to the set of MDs. We need a few notions before introducing it.

**Definition 9.** A set  $M$  of MDs is in *standard form* if no two MDs in  $M$  have the same expression to the left of the arrow. □

Notice that any set of MDs can be put in standard form by replacing subsets of MDs of the form  $\{R[\bar{A}] \approx S[\bar{B}] \rightarrow R[\bar{C}_1] \approx S[\bar{E}_1], \dots, R[\bar{A}] \approx S[\bar{B}] \rightarrow R[\bar{C}_n] \approx S[\bar{E}_n]\}$  by the single MD  $R[\bar{A}] \approx S[\bar{B}] \rightarrow R[\bar{C}] \approx S[\bar{E}]$ , where the set of corresponding pairs of attributes of  $(\bar{C}, \bar{E})$  is the union of

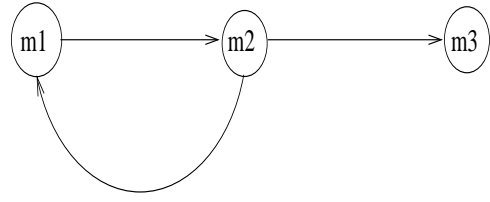


Figure 1: An MD-Graph

those of  $(\bar{C}_1, \bar{E}_1), \dots, (\bar{C}_n, \bar{E}_n)$ . From now on, we will assume that all sets of MDs are in standard form.

For an MD  $m$ ,  $\text{LHS}(m)$  and  $\text{RHS}(m)$  denote the sets of attributes that appear to the left side and to right side of the arrow, respectively.

**Definition 10.** Let  $M$  be a set of MDs in standard form. The *MD-graph* of  $M$ , denoted  $\text{MDG}(M)$ , is a directed graph with a vertex labeled  $m$  for each  $m \in M$ , and with an edge from  $m_1$  to  $m_2$  iff  $\text{RHS}(m_1) \cap \text{LHS}(m_2) \neq \emptyset$ . □

**Example 9.** Consider the set of MDs:  $m_1 : R[A] \approx S[B] \rightarrow R[C] \approx S[D]$ ,  $m_2 : R[C] \approx S[D] \rightarrow R[A] \approx S[B]$ ,  $m_3 : S[E] \approx S[B] \rightarrow T[F] \approx T[F]$ . It has the MD-graph shown in Figure 1. □

A set of MDs whose MD-graph contains edges is called *interacting*. Otherwise, it is *non-interacting*.

**Definition 11.** (a) A cycle  $C$  in an MD-graph  $\text{MDG}(M)$  is called a *simple cycle* if for each pair  $(m_1, m_2)$  of successive vertices in  $C$ , the corresponding pairs to the left of the arrow in  $m_2$  are corresponding pairs to the right of the arrow in  $m_1$ , and do not occur elsewhere in  $m_1$ . (b) A set  $M$  of MDs is *simple-cycle* if its MD-graph  $\text{MDG}(M)$  is a simple cycle. □

**Example 10.** The following is a simple-cycle set of MDs.

$$m_1 : R[A] \approx S[B] \rightarrow R[C, F] \approx S[E, G],$$

$$m_2 : R[C] \approx S[E] \wedge R[F] \approx S[G] \rightarrow R[A] \approx S[B].$$

The MD-graph is a cycle, because attributes in  $\text{RHS}(m_2)$  are in  $\text{LHS}(m_1)$ , and vice-versa. This cycle is a simple cycle, because the corresponding pairs  $(C, E)$  and  $(F, G)$  to the right of the arrow in  $m_1$  are corresponding pairs to the left of the arrow in  $m_2$ , and vice-versa. □

For this class of MDs it is easy to characterize the form an MRI takes. This is first illustrated with an example.

**Example 11.** Consider the instance  $D$  (with tuple ids) and simple-cycle set of MDs.

$R$	$A$	$B$	$C$
1	$a_1$	$d_1$	$f$
2	$a_2$	$e_2$	$g$
3	$b_1$	$e_1$	$h$
4	$b_2$	$d_2$	$i$

$$R[A] \approx R[A] \rightarrow R[B] \approx R[B],$$

$$R[B] \approx R[B] \rightarrow R[A] \approx R[A].$$

The only similarities are:  $a_i \approx a_j$ ,  $b_i \approx b_j$ ,  $d_i \approx d_j$ ,  $e_i \approx e_j$ , with  $i, j \in \{1, 2\}$ . If the MDs are applied twice, successively, to the instance, one possible result is:

	$A$	$B$	$C$			$A$	$B$	$C$
1	$a_1$	$d_1$	$f$	$\rightarrow$	1	$b_2$	$d_1$	$f$
2	$a_2$	$e_2$	$g$		2	$a_2$	$d_1$	$g$
3	$b_1$	$e_1$	$h$		3	$a_2$	$e_1$	$h$
4	$b_2$	$d_2$	$i$		4	$b_2$	$e_1$	$i$

	$A$	$B$	$C$			$A$	$B$	$C$
1	$a_2$	$e_1$	$f$	$\rightarrow$	1	$a_2$	$e_1$	$f$
2	$a_2$	$d_1$	$g$		2	$a_2$	$d_1$	$g$
3	$b_2$	$d_1$	$h$		3	$b_2$	$d_1$	$h$
4	$b_2$	$e_1$	$i$		4	$b_2$	$e_1$	$i$

From this it is clear that, in any sequence of states  $D, D_1, D_2, \dots$  obtained by applying the MDs, the updated instances must have the following pairs of values equal:

$D_i, i \text{ odd}$	Column	
	$A$	$B$
tuple (id) pairs	(1, 4), (2, 3)	(1, 2), (3, 4)
$D_i, i \text{ even}$	Column	
	$A$	$B$
tuple (id) pairs	(1, 2), (3, 4)	(1, 4), (2, 3)

In any stable instance, the pairs of values in the above tables must be equal. Clearly, this can only be the case if all values in the  $A$  and  $B$  columns are equal. This can be achieved with a single update, choosing any value as the common value. Thus, the MRIs of any instance are those with all values in the  $A$  and  $B$  columns set to their most common value. In the case of  $D$  above, there are 16 MRIs.  $\square$

The Algorithm *ComputeMRI* below generalizes the idea presented in Example 11. It computes the set of all MRIs for the case of an arbitrary simple cycle. (The relevant definitions are given below).

**Definition 12.** Let  $m$  be the MD  $R[\bar{A}] \approx S[\bar{B}] \rightarrow R[\bar{C}] \Leftarrow S[\bar{E}]$ . The transitive closure,  $T^\approx$ , of  $\approx$  is the transitive closure of the binary relation relation on tuples  $t_1[\bar{A}] \approx t_2[\bar{B}]$ , where  $t_1 \in R$  and  $t_2 \in S$ .  $\square$

Notice that Definition 12 implies that the transitive closure of  $\approx$  is an equivalence relation on the tuples of  $R$  and  $S$ . It therefore forms a partition of these tuples into disjoint equivalence classes.

**Definition 13.** For a set  $S$  of binary relations, the transitive closure,  $T^S$ , of  $S$  is the transitive closure of the union of all relations in  $S$ .  $\square$

This definition can be applied, in particular, to the  $T^\approx$ s in Definition 12, for several MDs. For the case in which  $S$  in Definition 13 is a set of equivalence relations,  $T^S$  is also an equivalence relation. These definitions are used in Algorithm *ComputeMRI* in Table 1.

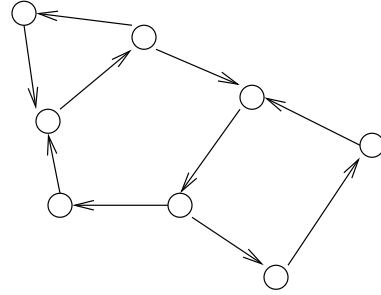
**Proposition 1.** Algorithm *ComputeMRI* returns the set of all MRIs of  $D$  wrt a simple-cycle set  $M$  of MDs.  $\square$

With some minor modifications to the  $T$  relation in Algorithm *ComputeMRI*, we can make the latter work also for sets of MDs whose vertices in the MD-graph can occur on more than one simple cycle, as shown in Figure 2.<sup>1</sup> The following HSC class of sets of MDs extends the simple-cycle class.

<sup>1</sup>The modification involves using the tuple-attribute closure introduced in Definition 17, for the cases where the MD-graph has more than one connected component.

**Table 1: Algorithm *ComputeMRI***

**Input:** A database instance  $D$  and a simple-cycle set  $M$  of MDs.  
**Output:** Set of MRIs of  $D$  with respect to  $M$ .  
1) **For**  $1 \leq j \leq n$   
2)     Compute  $T_j$ , the transitive closure of  $\approx_j$   
3)     Compute the transitive closure  $T$  of the set  $\{T_j | 1 \leq j \leq n\}$   
4)     **For** each corresponding pair of attributes  $(A, B)$  that appears in  $M$ :  
5)         **For** each equivalence class  $E$  defined by  $T$ :  
6)             Choose a value  $v$  from among the  $A$  and  $B$  attribute values of tuples in  $R \cap E$  and  $S \cap E$ , respectively, such that no other value occurs more frequently  
7)             **For** each tuple  $t \in E$ :  
8)                  $t[A] \leftarrow v$  if  $t \in R$   
9)                  $t[B] \leftarrow v$  if  $t \in S$   
10)     Repeat 4-9 for other choices of  $v$  to produce other MRIs  
11)     Return the resulting set of MRIs



**Figure 2: The MD-graph of an HSC set of MDs**

**Definition 14.** A set  $M$  of MDs is *hit simple cyclic* (HSC) iff each vertex in  $MDG(M)$  is on at least one simple cycle of  $MDG(M)$ .  $\square$

The next example shows that even for simple classes of MDs, there may be exponentially many MRIs.

**Example 12.** Consider the relational predicate  $R[A, B]$  and the MD  $m : R[A] \approx R[A] \rightarrow R[B] \Leftarrow R[B]$ . Let  $D$  be an instance of  $R$  with tuples  $\{t_i | 1 \leq i \leq n\}$ , for some even number  $n$ , such that: (a) The values in  $D$  satisfy the similarities  $t_i[A] \approx t_{i+1}[A]$  for all odd  $i$  with  $1 \leq i \leq n-1$ , and no others, and (b)  $t_i[B] \neq t_{i+1}[B]$  for all odd  $i$  with  $1 \leq i \leq n-1$ . It is clear that an MRI is obtained by setting the  $B$  attributes of  $t_i$  and  $t_{i+1}$  to either  $t_i[B]$  or  $t_{i+1}[B]$  for each odd  $i$  such that  $1 \leq i \leq n-1$ . The number of MRIs is the number of possible choices of such values, which is  $2^{n/2}$ .  $\square$

The MD in the previous example is HSC. Actually, the simple form of the MRIs for HSC sets can be used to obtain an upper bound for  $RA_{Q,M}$  that (under usual complexity-theoretic assumptions) is lower than exponential. This relies on the assumption that, if a resolved instance contains values outside the active domain of the original instance, then

those values are bounded above in length by a polynomial in the size of the original instance. This assumption is in accord with practical constraints on databases and any reasonable definition of similarity.

*Theorem 3.* For HSC sets of MDs, if resolved instances are restricted to contain values bounded in length by a polynomial in the length of the input, then problem  $RA_{Q,M}$  is in *coNP* for any first-order query  $Q$ .  $\square$

In this section, we established a complexity bound for  $RA_{Q,M}$  which holds for class of MDs with cyclic MD-graphs and all first-order queries. The bound follows from the simple form of the MRIs, as described by Algorithm *ComputeMRI*. In the next section, we further exploit this latter result to show that, for HSC sets and certain first-order queries, the resolved answers can be retrieved in polynomial time.

## 5. RESOLVED QUERY ANSWERING: TRACTABILITY AND REWRITING

In this section, we discuss tractable cases of  $RA_{Q,M}$ . In particular, we propose a query rewriting technique for obtaining the resolved answers for certain FO queries and MDs. In Section 6, we will relate  $RA_{Q,M}$  to *consistent query answering* (CQA) [7]. This connection and some known results in CQA will allow us to identify further tractable cases, but also to establish the intractability of  $RA_{Q,M}$  for certain classes of queries and MDs. The latter makes the tractability results obtained in this section even more relevant.

A possible approach to obtaining the resolved answers to a query  $Q$  from an instance  $D$  is to rewrite  $Q$  into a new query  $Q'$  on the basis of  $Q$  and  $M$ .  $Q'$  should be such that, when posed to  $D$  (as usual), it returns the resolved answers to  $Q$  from  $D$ . In this case, it is not necessary to explicitly compute the MRIs. If  $Q'$  can be efficiently evaluated against  $D$ , then the resolved answers can also be efficiently computed and  $RA_{Q,M}$  becomes tractable. This methodology was proposed in [3] for CQA.

This section investigates this query rewriting approach to the computation of resolved answers for HSC sets of MDs. The input queries  $Q$  will be conjunctive queries with certain restrictions on the joins. However, the rewritten queries  $Q'$  may involve aggregate operators (actually, *Count*), universal quantification, and Datalog rules (to specify the transitive closure). We will need to compute transitive closures and count the number of occurrences of values in order to enforce minimal change. In any case, the resulting query  $Q'$  will still be evaluable in polynomial time in the size of  $D$ .

Specifically, the input queries we consider have the form  $Q(\bar{x}) : \exists \bar{u}(R_1(\bar{v}_1) \wedge \dots \wedge R_n(\bar{v}_n))$ , where  $\bar{x} = (\cup \bar{v}_i) \setminus \bar{u}$ . For tractability of  $RA_{Q,M}$ , we need additional restrictions on them.

*Definition 15.* (a) For a set  $M$  of MDs defined on schema  $S$ , the *changeable attributes* of  $S$  are those that appear to the right of the arrow in some  $m \in M$ . The other attributes of  $S$  are called *unchangeable*.

(b) Let  $Q$  be a conjunctive query and  $M$  a set of MDs. Query  $Q$  is an *unchangeable attribute join* conjunctive query (*ucajCQ*) if there are no bound, repeated variables in  $Q$  that correspond to changeable attributes.  $\square$

*Example 13.* Let  $M$  be the single MD  $R[A] \approx R[A] \rightarrow R[B] \Leftarrow R[B]$ . The query  $Q(x, z) : \exists y(R(x, y) \wedge R(z, y))$

is not in the *ucajCQ* class, because it contains a bound, repeated variable ( $y$ ) which corresponds to a changeable attribute ( $B$ ). However, the query  $Q(y) : \exists x \exists z(R(x, y) \wedge R(x, z))$  is in *ucajCQ*, since the only bound, repeated variable ( $x$ ) corresponds to an unchangeable attribute ( $A$ ).  $\square$

In Section 6 we will encounter HSC MDs (even non-interacting MDs) and conjunctive queries outside *ucajCQ* for which  $RA_{Q,M}$  is intractable (cf. Theorem 5 below).

To incorporate counting into FO queries, we will use the operator *Count*( $R$ ) that returns the number of tuples in relation  $R$  (cf. [1]). *Count* will be applied to sets of tuples of the form  $\{\bar{t} \mid C\}$ , where  $\bar{t}$  is a tuple of variables, and  $C$  is a FO condition whose free variables include those in  $\bar{t}$ . Now we show a simple example of rewriting that uses *Count*.

*Example 14.* Consider a relation  $R$ , the MD  $R[A] \approx R[A] \rightarrow R[B] \Leftarrow R[B]$ , and the query  $Q(x, y, z) : R(x, y, z)$ .  $R$  and its (single) MRI are shown below.

$R$	$A$	$B$	$C$	MRI	$A$	$B$	$C$
	$a_1$	$b_1$	$c_1$		$a_1$	$b_2$	$c_1$
	$a_1$	$b_2$	$c_2$		$a_1$	$b_2$	$c_2$
	$a_1$	$b_2$	$c_3$		$a_1$	$b_2$	$c_3$

The set of resolved answers to  $Q$  is  $\{(a_1, b_2, c_1), (a_1, b_2, c_2), (a_1, b_2, c_3)\}$ . It is not difficult to see that the following query returns the resolved answers (for any initial instance of  $R$ ). In it,  $T$  stands for the transitive closure  $T^\approx$  of  $\approx$  (cf. Definition 12).

$$Q'(x, y, z) : \exists y' R(x, y', z) \wedge \forall y'' [Count\{(x', y, z') \mid T((x, y', z), (x', y, z')) \wedge R(x', y, z')\} > Count\{(x', y'', z') \mid T((x, y', z), (x', y'', z')) \wedge R(x', y'', z') \wedge y'' \neq y\}].$$

Intuitively, the first conjunct requires the existence of a tuple  $t$  with the same  $A$  and  $C$  attribute values as the answer. Since the values of these attributes are not changed when going from the original instance to an MRI, such a tuple must exist. However, the tuple is not required to have the same  $B$  attribute value as the answer tuple, because this attribute can be modified. For example,  $(a_1, b_2, c_1)$  is a resolved answer, but is not in  $R$ . What makes it a resolved answer is the fact that it is in an equivalence class of  $T$  (consisting of all three tuples in  $R$ ) for which  $b_2$  occurs more frequently as a  $B$  attribute value than any other value. This condition on resolved answers is expressed by the second conjunct.  $\square$

For simplicity, we present our query rewriting algorithm for non-interacting MDs, a special case of HSC sets of MDs where the connected components have only one vertex. The generalization to arbitrary HSC sets is straightforward, and the required modifications are indicated at the end of this section. First we require the following definitions.

*Definition 16.* Let  $M$  be a set of MDs on schema  $S$ . (a) Define a (symmetric) binary relation  $\Leftarrow_r$  which relates attributes  $R[A]$ ,  $S[B]$  of  $S$  if there is an MD in  $M$  where  $R[A] \Leftarrow S[B]$  appears to the right of the arrow.

(b) The *attribute closure*,  $T_{at}$ , of  $M$  is the binary relation on attributes defined as the reflexive, transitive closure of  $\Leftarrow_r$ .

(c) We use the notation  $E_{R[A]}$  to denote the equivalence class of  $T_{at}$  to which attribute  $R[A]$  belongs.  $\square$

Note that, in general, there will be pairs of attributes  $R[A]$ ,  $S[B]$  for which  $E_{R[A]} = E_{S[B]}$ .

*Example 15.* Let  $M$  be the set of MDs

$$\begin{aligned} R[A] &\approx_1 S[B] \rightarrow R[C] \Leftarrow S[D], \\ S[E] &\approx_2 T[F] \wedge S[G] \approx T[H] \rightarrow S[D, K] \Leftarrow T[J, L], \\ T[F] &\approx_3 T[H] \rightarrow T[L, N] \Leftarrow T[M, P]. \end{aligned}$$

The equivalence classes of  $T_{at}$  are  $E_{R[C]} = \{R[C], S[D], T[J]\}$ ,  $E_{S[K]} = \{S[K], T[L], T[M]\}$ , and  $E_{T[N]} = \{T[N], T[P]\}$ .  $\square$

To describe the MRIs in this case, we need the transitive closure relation defined below.

*Definition 17.* Let  $m$  be the MD  $R[\bar{A}] \approx S[\bar{B}] \rightarrow R[\bar{C}] \Leftarrow S[\bar{E}]$ .

(a) Let  $\approx'$  be the following binary relation on tuple-attribute pairs:  $(t_1, C) \approx' (t_2, E) :\Leftrightarrow t_1[\bar{A}] \approx t_2[\bar{B}]$  and  $(C, E)$  is a corresponding pair of  $(\bar{C}, \bar{E})$ .

(b) The *tuple-attribute closure*  $TA$  of  $m$  is the reflexive, transitive closure of  $\approx'$ .  $\square$

We denote by  $TS$  the transitive closure of a set of tuple-attribute closures (cf. Definition 13).  $TS$  partitions the set of tuple/attribute pairs into disjoint equivalence classes.

To keep the notation simple, we omit parentheses delimiting tuples and tuple/attribute pairs when writing the arguments of  $TA$  and  $TS$ . For example, for tuples  $t_2 = (a, b, c)$  and  $t_3 = (d, e, f)$  with attributes  $A$  and  $C$ , respectively,  $TS(((a, b, c), A), ((d, e, f), C))$  is written as  $TS(a, b, c, A, d, e, f, C)$ .

Algorithm *Rewrite* in Table 2, outputs a rewritten query  $\mathcal{Q}'$  that returns the resolved answers to a given input conjunctive query  $\mathcal{Q}$  and set of non-interacting MDs. This is done by separately rewriting each conjunct  $R_i(\bar{v}_i)$  in  $\mathcal{Q}$ . If  $R_i(\bar{v}_i)$  contains no free variables, then it is unchanged (line 5). Otherwise, it is replaced with a conjunction involving the same atom and additional conjuncts which use the *Count* operator. The conjuncts involving *Count* express the condition that, for each changeable attribute value returned by the query, this value is more numerous than any other value in the same set of values that is equated by the MDs. The *Count* expressions contain new local variables as well as a new universally quantified variable  $v'_{iA}$ .

*Example 16.* We illustrate the algorithm with predicates  $R[ABC], S[EFG], U[HI]$ , the query  $\mathcal{Q}(x, y, z) : \exists t, u, p, q (R(x, y, z) \wedge S(t, u, z) \wedge U(p, q))$ ; and the MDs:  $R[A] \approx S[E] \rightarrow R[B] \Leftarrow S[F]$  and  $S[E] \approx U[H] \rightarrow S[F] \Leftarrow U[I]$ .

Since the  $S$  and  $U$  atoms have no free variables holding the values of changeable attributes, these conjuncts remain unchanged (line 5). The only free variable holding the value of a changeable attribute is  $y$ . Therefore, line 7 sets  $\bar{v}'_1$  to  $(x, y', z)$ . Variable  $y$  contains the value of attribute  $R[B]$ . The equivalence class  $E_{R[B]}$  of  $T_{at}$  to which  $R[B]$  belongs is  $\{R[B], S[F], U[I]\}$ , so the loop at line 11 generates the atoms  $R(x', y, z')$ ,  $R(x', y'', z')$ ,  $S(t', y, z')$ ,  $S(t', y'', z')$ ,  $U(p', y)$ ,  $U(p', y'')$ . The rewritten query is obtained by replacing in  $\mathcal{Q}$  the conjunct  $R(x, y, z)$  by  $\exists y'(R(x, y', z) \wedge \forall y''[$

$$\begin{aligned} &Count\{(x', y, z') | TS(x, y', z, R[B], x', y, z', R[B]) \wedge \\ &R(x', y, z')\} + Count\{(t', y, z') | TS(x, y', z, R[B], \\ &t', y, z', S[F]) \wedge S(t', y, z')\} + Count\{(p', y) | TS(x, y', z, \\ &R[B], p', y, U[I]) \wedge U(p', y)\} > \end{aligned}$$

**Table 2: Algorithm *Rewrite***

<b>Input:</b>	A query in <i>ucajCQ</i> and non-interacting set of MDs $M$ .
<b>Output:</b>	The rewritten query $\mathcal{Q}'$ .
1)	Let $\mathcal{Q}(\bar{t}) : \exists \bar{u} \wedge_{1 \leq i \leq n} R_i(\bar{v}_i)$ be the query.
2)	<b>For</b> each $R_i(\bar{v}_i)$
3)	Let $C$ be the set of changeable attributes of $R_i$ corresponding to a free variable in $\bar{v}_i$
4)	<b>If</b> $C$ is empty
5)	$Q_i(\bar{v}_i) \leftarrow R_i(\bar{v}_i)$
6)	<b>Else</b>
7)	Let $\bar{v}'_i$ be $\bar{v}_i$ with each variable $v_{iA}$ in $\bar{v}_i$ holding the value of an attribute $A \in C$ replaced by a new variable $v'_{iA}$
8)	Let $\bar{v}_{iC}$ be the vector of variables $v_{iA}$ , $A \in C$
9)	Let $\bar{v}'_{iC}$ be the vector of variables $v'_{iA}$ , $A \in C$
10)	<b>For</b> each variable $v_{iA}$ in $\bar{v}_{iC}$
11)	<b>For each attribute</b> $R_j[B_k] \in E_A$
12)	Generate atom $R_j(\bar{u}_{jk})$ , where all variables in $\bar{u}_{jk}$ are new except the one holding the value of $R_j[B_k]$ , which is $v_{iA}$
13)	Generate atom $R_j(\bar{w}_{jk})$ , where all variables in $\bar{w}_{jk}$ are labelled as in $\bar{u}_{jk}$ except the one holding the value of $R_j[B_k]$ , which is $v'_{iA}$
14)	$C_{jk}^{A1} \leftarrow Count\{\bar{u}_{jk}   TS(\bar{v}'_i, R_i[A], \bar{u}_{jk}, R_j[B_k]) \wedge R_j(\bar{u}_{jk})\}$
15)	$C_{jk}^{A2} \leftarrow Count\{\bar{w}_{jk}   TS(\bar{v}'_i, R_i[A], \bar{w}_{jk}, R_j[B_k]) \wedge R_j(\bar{w}_{jk}) \wedge v'_{iA} \neq v_{iA}\}$
16)	$Q_i(\bar{v}_i) \leftarrow \exists \bar{v}'_{iC} \{R_i(\bar{v}'_i) \wedge A \in C \forall v'_{iA} [\Sigma_{j,k} C_{jk}^{A1} > \Sigma_{j,k} C_{jk}^{A2}]\}$
17)	$\mathcal{Q}'(\bar{t}) \leftarrow \exists \bar{u} \wedge_{1 \leq i \leq n} Q_i(\bar{v}_i)$
18)	<b>return</b> $\mathcal{Q}'$

$$\begin{aligned} &Count\{(x', y'', z') | \\ &TS(x, y', z, R[B], x', y'', z', R[B]) \wedge R(x', y'', z') \wedge \\ &y'' \neq y\} + Count\{(t', y'', z') | TS(x, y', z, R[B], t', y'', \\ &z', S[F]) \wedge S(t', y'', z') \wedge y'' \neq y\} + Count\{(p', y'') | \\ &TS(x, y', z, R[B], p', y'', U[I]) \wedge U(p', y'') \wedge y'' \neq y\}. \quad \square \end{aligned}$$

*Theorem 4.* For a set  $M$  of non-interacting MDs and a query  $\mathcal{Q}$  in the class *ucajCQ*, the query  $\mathcal{Q}'$  computed by Algorithm *Rewrite* returns the resolved answers to  $\mathcal{Q}$  when posed to any instance.  $\square$

As expected, the rewriting algorithm that produced the rewritten query does not depend upon the dirty instance at hand, but only on the MDs and the input query, and runs in polynomial time.

Algorithm *Rewrite* can be easily adapted and extended to handle HSC sets of MDs. All that is required is a modification to the tuple-attribute closure in Definition 17, as follows: For an HSC set of MDs  $M$  and  $m \in M$ , a pair of tuples  $t_1$  and  $t_2$  satisfies  $(t_1, C) \approx' (t_2, E)$  iff  $t_1[\bar{A}] \approx t_2[\bar{B}]$



and  $(C, E)$  appears as a corresponding pair to the right of the arrow in some MD in the same connected component of the MD graph as  $m$ . Tuple-attribute closure is redefined as the transitive closure of this new relation. As with Theorem 4, the correctness proof is based on the simple form of the MRIs, and is proved using the same technique as in the proof of Proposition 1.

## 6. THE CQA CONNECTION

MDs can be seen as a new form of integrity constraint (IC). An instance  $D$  violates an MD  $m$  if there are unresolved duplicates, i.e. tuples  $t_1$  and  $t_2$  in  $D$  that satisfy the similarity condition of  $m$ , but differ on some pair of attributes that are matched by  $m$ . The instances that are consistent with a set of MDs  $M$  are resolved instances of themselves with respect to  $M$ . Among classical ICs, the closest analogues of MDs are functional dependencies (FDs).

Given a database instance  $D$  and a set of ICs  $\Sigma$ , possibly not satisfied by  $D$ , consistent query answering (CQA) is the problem of characterizing and computing the answers to queries  $Q$  that are true in all the instances  $D'$  that are consistent with  $\Sigma$  and minimally differ from  $D$  [3]. The consistent instances  $D'$  are called *repairs*. Minimal difference can be defined in different ways. Most of the research in CQA has concentrated on the case where the symmetric difference of instances, as sets of tuples, is made minimal under set inclusion [3, 7, 11]. However, also the minimization of the cardinality of this difference has been investigated [20, 2]. Other forms of minimization measure the differences in attribute values between  $D$  and  $D'$  [17, 21, 16, 8]. Because of their practical importance, much work on CQA has been done for the case where  $\Sigma$  is a set of functional dependencies (FDs), in particular, key constraints (KCs) [12, 18, 23, 22, 24].

Actually, for a set of KCs  $\mathcal{K}$  and repairs based on tuple deletions, a *repair*  $D'$  of an instance  $D$  can be characterized as a maximal subset of  $D$  that satisfies  $\mathcal{K}$ :  $D' \subseteq D$ ,  $D' \models \mathcal{K}$  and there is no  $D''$  with  $D' \subsetneq D'' \subseteq D$ , with  $D'' \models \mathcal{K}$  [12].

Now, for a FO query  $Q(\bar{x})$  and a set of KCs  $\mathcal{K}$ , the *consistent query answering problem* is about deciding membership of the set

$$CQA_{Q, \mathcal{K}} = \{(D, \bar{a}) \mid \bar{a} \text{ is an answer to } Q \text{ in all repairs of } D \text{ with respect to } \mathcal{K}\}.$$

A  $\bar{a}$  satisfying the above is called a *consistent answer* to  $Q$  from  $D$ .

Notice that this notion of minimality involved in repairs wrt FDs is tuple and set-inclusion oriented, whereas the one related to MRIs (cf. Definition 6) is attribute and cardinality oriented. However, the connection can still be established. In particular, the following result can be obtained from [12, Thm. 3.3].

*Theorem 5.* Consider a relational predicate  $R[A, B, C]$ , the MD

$$m: R[A] = R[A] \rightarrow R[B, C] \rightleftharpoons R[B, C], \quad (5)$$

and the query  $Q: \exists x \exists y \exists y' \exists z (R(x, y, c) \wedge R(z, y', d) \wedge y = y')$ .  $RA_{Q, \{m\}}$  is *coNP*-complete.  $\square$

Notice that the conjunctive query in this result does not belong to the *ucjCQ* class.

For certain classes of conjunctive queries and ICs consisting of a single KC per relation, CQA has been proved to be tractable. This is the case for the  $\mathcal{C}_{forest}$  class of conjunctive queries [18]. Actually, for this class there is a FO rewriting of the original query that returns the certain answers.  $\mathcal{C}_{forest}$  excludes repeated relations and allows joins only between non-key and key attributes. Similar results were subsequently proved for a larger class of queries that includes some queries with repeated relations and joins between non-key attributes [23, 22, 24]. The following result allows us to take advantage of tractability results for CQA in our MD setting.

*Proposition 2.* Let  $D$  be a database instance with a single relation  $R$ . Let  $m$  be a MD of the form  $R[\bar{A}] = R[\bar{A}] \rightarrow R[\bar{B}] \rightleftharpoons R[\bar{B}]$ , where the set of attributes of  $R$  is  $\bar{A} \cup \bar{B}$  and  $\bar{A} \cap \bar{B} = \emptyset$ . Then there is a polynomial time reduction from  $RA_{Q, \{m\}}$  to  $CQA_{Q, \{\kappa\}}$ , where  $\kappa$  is the key constraint  $\bar{A} \rightarrow \bar{B}$ .  $\square$

Proposition 2 can be easily generalized to several relations with one such MD defined on each. The reduction takes an instance  $D$  for  $RA_{Q, \{m\}}$  and produces an instance  $D'$  for  $CQA_{Q, \{\kappa\}}$ . The schema of  $D'$  is the same for  $D$ , but the extensions of the relational predicates in it are changed wrt  $D$  via counting. Since definitions for those aggregations can be included (or inserted) in the query  $Q$ , we obtain:

*Theorem 6.* Let  $\mathcal{S}$  be a database schema with relation predicates  $R_i$ ,  $1 \leq i \leq n$  with a set  $\mathcal{K}$  of KCs  $\kappa_i: R_i[\bar{A}_i] \rightarrow R_i[\bar{B}_i]$ ,  $1 \leq i \leq n$ . Let  $Q$  be a FO query, and suppose there exists a polynomial time computable FO query  $Q'$ , such that  $Q'$  returns the consistent answers to  $Q$  from  $D$ . Then there exists a polynomial time computable FO query  $Q''$  with aggregation that returns the resolved answers to  $Q$  from  $D$  wrt the MDs  $m_i: R_i[\bar{A}_i] = R_i[\bar{A}_i] \rightarrow R_i[\bar{B}_i] \rightleftharpoons R_i[\bar{B}_i]$ ,  $1 \leq i \leq n$ .  $\square$

The aggregation in  $Q''$  in Theorem 6 arises from the transformation of the instance that is used in the reduction in Proposition 2. We emphasize that  $Q''$  is *not* obtained using algorithm *Rewrite* from Section 5, which is not guaranteed to work for queries outside the class *ucjCQ*. Rather, a first-order transformation of the  $R_i$  relations with *Count* is composed with  $Q'$  to produce  $Q''$ . Similar to Algorithm *Rewrite* in Section 5, they are used to express the most frequently occurring values for the changeable attributes for a given set of tuples with identical values for the unchangeable attributes.

This theorem can be applied to decide/compute resolved answers through composition in those cases where a FO rewriting for CQA has been identified. In consequence, it extends the tractable cases identified in Section 5. They can be applied to queries that are not in *ucjCQ*.

*Example 17.* The query  $Q: \exists x \exists y \exists z (R(x, y) \wedge S(y, z))$  is in the class  $\mathcal{C}_{forest}$  for relational predicates  $R[A, B]$  and  $S[C, E]$  and FDs  $A \rightarrow B$  and  $C \rightarrow E$ . By Theorem 6 and the results in [18], this implies the existence of a polynomial time computable FO query with counting that returns the resolved answers to  $Q$  wrt MDs  $R[A] = R[A] \rightarrow R[B] \rightleftharpoons R[B]$  and  $S[C] = S[C] \rightarrow S[E] \rightleftharpoons S[E]$ . Notice that  $Q$  is not in *ucjCQ*, since the bound variable  $y$  is associated with the changeable attribute  $R[B]$ .  $\square$

## 7. CONCLUSIONS

In this paper we have proposed a revised semantics for matching dependency (MD) satisfaction wrt the one originally proposed in [15]. The main outcomes from that semantics are the notions of *minimally resolved instance* (MRI) and *resolved answers* (RAs) to queries. The former capture the intended, clean instances obtained after enforcing the MDs on a given instance. The latter are query answers that persist across all the MRIs, and can be considered as robust and semantically correct answers.

We investigated the new semantics, the MRIs and the RAs. We considered the existence of MRIs, their number, and the cost of computing them. Depending on syntactic criteria on MDs and queries, tractable and intractable cases of resolved query answering were identified. The tractable cases coincide with those where the original query can be rewritten into a new, polynomial-time evaluable query that returns the resolved answers when posed to the original instance. It is interesting that the rewritings make use of counting and recursion (for the transitive closure). The original queries considered in this paper are all conjunctive. Other classes of queries will be considered in future work.

Many of our results apply to cases for which the resolved instances can be obtained after a single (batch) update operation. The investigation of cases requiring multiple updates is a subject of ongoing research. We have obtained several tractability and intractability results. However, understanding the complexity landscape requires still much more research.

We established interesting connections between resolved query answering wrt MDs and consistent query answers. There are still many issues to explore in this direction, e.g. the possible use of logic programs with stable model semantics to specify the MRIs, so as it has been done with database repairs [4, 5, 19].

We have proposed some efficient algorithms for resolved query answering. Implementing them and experimentation are also left for future work. Notice that those algorithms use different forms of transitive closure. To avoid unacceptably slow query processing, it may be necessary to compute transitive closures off-line and store them. The use of Datalog with aggregate functions should also be investigated in this direction.

In this paper we have not considered cases where the matchings of attribute values, whenever prescribed by the MDs' conditions, are made according to matching functions. This element adds an entirely new dimension to the semantics and the problems investigated here. It certainly deserves investigation.

## 8. REFERENCES

- [1] S. Abiteboul, R. Hull, and V. Vianu. *Foundations of Databases*. Addison-Wesley, Don Mills, Ontario, 1995.
- [2] F. Afrati and P. Kolaitis. Repair checking in inconsistent databases: Algorithms and complexity. In *Proc. ICDT*, pages 31–41, 2009.
- [3] M. Arenas, L. Bertossi, and J. Chomicki. Consistent query answers in inconsistent databases. In *Proc. PODS*, pages 68–79, 1999.
- [4] M. Arenas, L. Bertossi, and J. Chomicki. Answer sets for consistent query answering in inconsistent databases. *Theory and Practice of Logic Programming*, 3(4-5):393–424, 2003.
- [5] P. Barceló, L. Bertossi, and L. Bravo. Characterizing and computing semantically correct answers from databases with annotated logic and answer sets. In *Semantics in Databases*, pages 7–33, 2003.
- [6] O. Benjelloun, H. Garcia-Molina, D. Menestrina, Q. Su, S. Euijong Whang, and J. Widom. Swoosh: A generic approach to entity resolution. *VLDB Journal*, 18(1):255–276, 2009.
- [7] L. Bertossi. Consistent query answering in databases. *ACM Sigmod Record*, 35(2):68–76, 2006.
- [8] L. Bertossi, L. Bravo, E. Franconi, and A. Lopatenko. The complexity and approximation of fixing numerical attributes in databases under integrity constraints. *Information Systems*, 33(4):407–434, 2008.
- [9] L. Bertossi and J. Chomicki. Query answering in inconsistent databases. In *Logics for Emerging Applications of Databases*, pages 43–83. Springer, 2003.
- [10] J. Bleiholder and F. Naumann. Data fusion. *ACM Computing Surveys*, 41(1):1–41, 2008.
- [11] J. Chomicki. Consistent query answering: Five easy pieces. In *Proc. ICDT*, pages 1–17, 2007.
- [12] J. Chomicki and J. Marcinkowski. Minimal-change integrity maintenance using tuple deletions. *Information and Computation*, 197(1/2):90–121, 2005.
- [13] A. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 19(1):1–16, 2007.
- [14] J. Fan. Dependencies revisited for improving data quality. In *Proc. PODS*, pages 159–170, 2008.
- [15] J. Fan, X. Jia, J. Li, and S. Ma. Reasoning about record matching rules. In *Proc. VLDB*, pages 407–418, 2009.
- [16] S. Flesca, F. Furfaro, and F. Parisi. Consistent query answers on numerical databases under aggregate constraints. In *Proc. DBPL*, pages 279–294, 2005.
- [17] E. Franconi, A. Laureti Palma, N. Leone, S. Perri, and F. Scarcello. Census data repair: A challenging application of disjunctive logic programming. In *Proc. LPAR*, pages 561–578, 2001.
- [18] A. Fuxman and R. Miller. First-order query rewriting for inconsistent databases. *J. Computer and System Sciences*, 73(4):610–635, 2007.
- [19] G. Greco, S. Greco, and E. Zumpano. A logical framework for querying and repairing inconsistent databases. *IEEE Transactions on Knowledge and Data Engineering*, 15(6):1389–1408, 2003.
- [20] A. Lopatenko and L. Bertossi. Complexity of consistent query answering in databases under cardinality-based and incremental repair semantics. In *Proc. ICDT*, pages 179–193, 2007.
- [21] J. Wijsen. Database repairing using updates. *ACM Transactions on Database Systems*, 30(3):722–768, 2005.
- [22] J. Wijsen. Consistent query answering under primary keys: A characterization of tractable cases. In *Proc. ICDT*, pages 42–52, 2009.
- [23] J. Wijsen. On the consistent rewriting of conjunctive queries under primary key constraints. *Information*

- [24] J. Wijsen. On the first-order expressibility of computing certain answers to conjunctive queries over uncertain databases. In *Proc. PODS*, pages 179–190, 2010.

## APPENDIX

### A. AUXILIARY RESULTS AND PROOFS

**Proof of Theorem 1:** Consider an undirected graph  $G$  whose vertices are labelled by pairs  $(t, A)$ , where  $t$  is a tuple identifier and  $A$  is an attribute of  $t$ . There is an edge between two vertices  $(s, A)$  and  $(t, B)$  iff  $s$  and  $t$  satisfy the similarity condition of some MD  $m \in M$  such that  $A$  and  $B$  are matched by  $m$ .

Update  $D$  as follows. Choose a vertex  $(t_1, A)$  such that there is another vertex  $(t_2, B)$  connected to  $(t_1, A)$  by an edge and  $t_1[A]$  and  $t_2[B]$  must be made equal to satisfy the equalities in condition 1. of Definition 3. For convenience in this proof, we say that  $t_2$  is unequal to  $t_1$  for such a pair of tuples  $t_1$  and  $t_2$ . Perform a breadth first search (BFS) on  $G$  starting with  $(t_1, A)$  as level 0. During the search, if a tuple is discovered at level  $i + 1$  that is unequal to an adjacent tuple at level  $i$ , the value of the attribute in the former tuple is modified so that it matches that of the latter tuple. When the BFS has completed, another vertex with an adjacent unequal tuple is chosen and another BFS is performed. This continues until no such vertices remain. It is clear that the resulting updated instance  $D'$  satisfies condition 1. of definition 3.

We now show by induction on the levels of the breadth first searches that for all vertices  $(t, A)$  visited,  $t[A]$  is modifiable. This is true in the base case, by choice of the starting vertex. Suppose it is true for all levels up to and including the  $i^{th}$  level. By definition of the graph  $G$  and condition 2. of definition 2, the statement is true for all vertices at the  $(i + 1)^{th}$  level. This proves the first statement of the theorem.

To prove the second statement, we show that, to satisfy condition 1. of Definition 3, the attribute values represented by each vertex in each connected component of  $G$  must be changed to a common value in the new instance. The statement then follows from the fact that the update algorithm can be modified so that the attribute value for the initial vertex in each BFS is updated to some arbitrary value at the start (since it is modifiable). By condition 1. of Definition 3, the pairs of values that must be equal in the updated instance  $D'$  correspond to those vertices that are connected by an edge in  $G$ . This fact and transitivity of equality imply that all attribute values in a connected component must be updated to a common value.  $\square$

**Proof of Theorem 2:** We give an algorithm to compute a resolved instance, and use a monotonicity property to show that it always terminates. For attribute domain  $d$  in  $D$ , consider the set  $S^d$  of pairs  $(t, A)$  such that attribute  $A$  of the tuple with identifier  $t$  has domain  $d$ . Let  $\{S_1, S_2, \dots, S_n\}$  be a partition of  $S^d$  into sets such that all tuple/attribute pairs in a set have the same value in  $D$ . Define the level of  $(t, A)$  to mean  $|S_j|$  where  $(t, A) \in S_j$ .

The algorithm first applies all MDs in  $M$  to  $D$  by setting equal pairs of unequal values according to the MDs. Specifically, consider a connected component  $C$  of the graph in the proof of Theorem 1. If the values of  $t[A]$  for all pairs  $(t, A)$

in  $C$  are not all the same, then their values are modified to a common value which is that of the pair with the highest level. This update is allowed by Theorem 1. In the case of a tie, the common value is chosen as the largest of the values according to some total ordering of the values from the domain that occur in the instance. It is easily verified that this operation increases the sum over all the levels of the elements of  $S^d$ , where  $d$  is the domain of the attributes of the pairs in  $C$ . These updates produce an instance  $D_1$  such that  $(D, D_1) \models M$ .

The MDs of  $M$  are then applied to the instance  $D_1$  to obtain a new instance  $D_2$  such that  $(D_1, D_2) \models M$  and so on, until a stable instance is reached. For each new instance, the sum over all domains  $d$  of the levels of the  $(t, A) \in S^d$  is greater than for the previous instance. Since this quantity is bounded above, the algorithm terminates with a resolved instance.  $\square$

For the proof of Proposition 1, we need an auxiliary result.

**Lemma 1.** Let  $D$  be an instance and let  $m$  be the MD in Definition 12. Let  $T$  be the transitive closure of  $\approx$ . An instance  $D'$  obtained by changing modifiable attribute values of  $D$  satisfies  $(D, D') \models m$  iff for each equivalence class of  $T$ , there is a constant vector  $\bar{v}$  such that, for all tuples  $t$  in the equivalence class,

$$\begin{aligned} t'[\bar{C}] &= \bar{v} \text{ if } t \in R(D) \\ t'[\bar{E}] &= \bar{v} \text{ if } t \in S(D) \end{aligned}$$

where  $t'$  is the tuple in  $D'$  with the same identifier as  $t$ .

*Proof:* Suppose  $(D, D') \models m$ . By Definition 3, for each pair of tuples  $t_1 \in R(D)$  and  $t_2 \in S(D)$  such that  $t_1[\bar{A}] \approx t_2[\bar{B}]$ ,

$$t'_1[\bar{C}] = t'_2[\bar{E}]$$

Therefore, if  $T(\bar{t}_1, \bar{t}_2)$  is true, then  $t'_1$  and  $t'_2$  must be in the transitive closure of the binary relation expressed by  $t'_1[\bar{C}] = t'_2[\bar{E}]$ . But the transitive closure of this relation is the relation itself (because of the transitivity of equality). Therefore,  $t'_1[\bar{C}] = t'_2[\bar{E}]$ . The converse is trivial.  $\square$

**Proof of Proposition 1:** Consider an input  $D, M$  to *ComputeMRI* with  $M$  a simple-cycle set of MDs given by

$$\begin{aligned} R[\bar{A}_0] \approx_0 S[\bar{B}_0] &\rightarrow R[\bar{A}'_0] \Leftarrow S[\bar{B}'_0] \\ R[\bar{A}_1] \approx_1 S[\bar{B}_1] &\rightarrow R[\bar{A}'_1] \Leftarrow S[\bar{B}'_1] \\ &\vdots \\ R[\bar{A}_{n-1}] \approx_{n-1} S[\bar{B}_{n-1}] &\rightarrow R[\bar{A}'_{n-1}] \Leftarrow S[\bar{B}'_{n-1}] \end{aligned}$$

Let  $T_j$  denote the transitive closure of the relation  $\approx_j$ . Let  $D_i$  denote an instance obtained by updating  $D$   $i$  times according to  $M$ , and for a tuple  $t \in D$ , denote the tuple with the same identifier in  $D_i$  by  $t^i$ . By Lemma 1 and straightforward induction, it can be seen that, after  $D$  has been updated  $i$  times,  $i \geq 1$ <sup>2</sup> according to  $M$  to obtain an instance  $D_i$ , for all tuples  $t$  in a given equivalence class  $E$  of  $T_j$ ,

$$t^i[\bar{A}'_{(j+i-1) \bmod n}] = \bar{v}_{ij}^E \text{ if } t \in R(D) \quad (6)$$

$$t^i[\bar{B}'_{(j+i-1) \bmod n}] = \bar{v}_{ij}^E \text{ if } t \in S(D) \quad (7)$$

<sup>2</sup>We use the term “update” even if a resolved instance is obtained after fewer than  $i$  modifications. In this case, the “update” is the identity mapping on all values.

for some vector of values  $\bar{v}_{ij}^E$ . Let  $D'$  be a resolved instance.  $D'$  satisfies the property that any number of applications of the MDs does not change the instance. Therefore,  $D'$  must satisfy (6) and (7) for all  $i$ . That is, for any  $T_j$ ,  $1 \leq j \leq n$ , for any equivalence class of  $T_j$ , for all tuples  $t$  in the equivalence class, and for  $1 \leq i \leq n$ ,

$$t'[\bar{A}'_i] = \bar{v}_{ij}^E \quad \text{if } t \in R(D) \quad (8)$$

$$t'[\bar{B}'_i] = \bar{v}_{ij}^E \quad \text{if } t \in S(D) \quad (9)$$

for some vector of values  $\bar{v}_{ij}^E$ , where  $t'$  is the tuple in  $D'$  with the same identifier as  $t$ .

Let  $T$  be the transitive closure of the set  $\{T_j | 1 \leq j \leq n\}$  (cf. definition 13). By (8) and (9), for any pair of tuples  $t_1$  and  $t_2$  satisfying  $T(t_1, t_2)$ ,  $t'_1$  and  $t'_2$  must satisfy  $T'(t'_1, t'_2)$ , where  $T'$  is the transitive closure of the binary relation on tuples expressed by  $t'_1[\bar{A}'_i] = t'_2[\bar{B}'_i]$ ,  $1 \leq i \leq n$ . Since the equality relation is closed under transitive closure, this implies the following property:

$$T(t_1, t_2) \text{ implies } t'_1[\bar{A}'_i] = t'_2[\bar{B}'_i], \quad 1 \leq i \leq n \quad (10)$$

It remains to show that the instances produced by *ComputeMRI* are resolved instances. That they are the MRIs will then follow from the fact that they have the fewest changes among all instances satisfying (10). For any equivalence class  $E$  of  $T$ , let  $\bar{v}_i^E$  be a list of values chosen by *ComputeMRI* as the common values for the pair of attribute lists  $(\bar{A}'_i, \bar{B}'_i)$  for tuples in  $E$ . To obtain the instance output by *ComputeMRI* for this choice of values,  $D$  can be updated as follows. For the  $i^{\text{th}}$  update, if the values of the attributes  $\bar{A}'_i$  and  $\bar{B}'_i$  must be modified to achieve (6) and (7), take  $\bar{v}_{ij}^E = \bar{v}_i^{E'}$ , where  $E'$  is the equivalence class of  $T$  that contains the equivalence class  $E$  of  $T_j$ . Note that such an  $E'$  always exists, and the assignment of values is consistent since overlapping equivalence classes  $T_i$  and  $T_j$  will be contained in the same equivalence class of  $T$ . Then after  $n$  updates, the resulting instance satisfies (10), with common values as chosen by *ComputeMRI*.

We must show that the resolved instance produced by this update process is the same instance that *ComputeMRI* returns for the given choice of update values. For any intermediate instance  $I$  obtained in this update process, let  $t_I$  denote the tuple in  $I$  with the same identifier as  $t$ . We will show by induction on the number of updates that were made to obtain  $I$  that for any  $i$ , whenever  $T_i(t_I, t'_I)$  for tuples  $t$  and  $t'$ , it holds that  $T(t, t')$ . This implies that updates made to  $t[A]$  for tuple  $t$  and attribute  $A$  can only set it equal to the common value for the equivalence class of  $T$  to which  $t$  belongs. Since *ComputeMRI* also sets  $t[A]$  to this value, this will prove the theorem.

By definition of  $T$ , if 0 updates were used to obtain  $I$ ,  $T_i(t_I, t'_I)$  implies  $T_i(t, t')$  implies  $T(t, t')$ . Assume it is true for instances obtained after at most  $k$  updates. Let  $I$  be an instance obtained after  $k+1$  updates. Suppose for the sake of contradiction that there exist tuples  $t_I$  and  $t'_I$  such that for some  $i$ ,  $T_i(t_I, t'_I)$  but  $\neg T(t, t')$ . Since  $\neg T(t, t')$  implies  $\neg T_i(t, t')$ , at least one of  $t[\bar{A}'_i]$  and  $t'[\bar{B}'_i]$  was updated so that  $T_i(t_I, t'_I)$ . We will assume that only  $t[\bar{A}'_i]$  was updated. The other cases are similar. Then it must have been updated to  $t''[\bar{A}'_i]$  or  $t''[\bar{B}'_i]$  for some  $t'' \in R$  or  $t'' \in S$ , respectively, such that, for the instance  $I'$  on which the update was performed, it holds that  $T_i(t_I, t'_I)$  and  $T_i(t'_I, t''_I)$ . By the induction hypothesis,  $T(t, t'')$  and  $T(t', t'')$ , which by the transitivity

of  $T$  implies  $T(t, t')$ , a contradiction.  $\square$

**Proof of Theorem 3:** If it can be verified in polynomial time that an instance is an MRI of a given instance wrt a set  $M$  of MDs, then  $RA_{Q,M}$  is in co-NP for any FO  $Q$ . This is because, for a given instance  $(D, t)$  of  $RA_{Q,M}$ ,  $t$  can be shown not to be a certain answer by guessing an instance  $D'$ , verifying that it is an MRI, and verifying that  $t$  is not an answer to  $Q$  for  $D'$ . Algorithm *ComputeMRI* can easily be modified to produce such a polynomial time verifier: compute the transitive closure relation  $T$  but instead of setting values equal, check that they are equal in the candidate MRI.  $\square$

*Lemma 2.* Let  $M$  be a non-interacting set of MDs of the form

$$\begin{aligned} m_1 : & \quad R_1[A_1] \approx_1 R_2[B_1] \rightarrow R_1[\bar{A}_2] = R_2[\bar{B}_2] \\ m_2 : & \quad R_3[A_3] \approx_2 R_4[B_3] \rightarrow R_3[\bar{A}_4] = R_4[\bar{B}_4] \\ & \quad \vdots \\ m_n : & \quad R_{2n-1}[A_{2n-1}] \approx_n R_{2n}[B_{2n-1}] \rightarrow \\ & \quad R_{2n-1}[\bar{A}_{2n}] = R_{2n}[\bar{B}_{2n}] \end{aligned}$$

Let  $T$  be the transitive closure of the set  $\{TA_1, TA_2, \dots, TA_n\}$ , where  $TA_i$  is the tuple-attribute closure of  $m_i$ . Then, for any instance  $D$ , an instance  $D'$  obtained by updating modifiable values of  $D$  is a resolved instance of  $D$  iff whenever  $T(t_1, A, t_2, B)$ ,  $t'_1[A] = t'_2[B]$ , where  $t'$  is the tuple in  $D'$  with the same identifier as  $t$ .

*Proof:* Suppose  $D'$  is a resolved instance. Since  $M$  is non-interacting, this implies  $(D, D') \models M$ . It is a corollary of Lemma 1 that whenever  $T(t_1, A, t_2, B)$ ,  $t'_1[A] = t'_2[B]$ , for all  $1 \leq i \leq n$ . The converse follows from the fact that, whenever a pair of tuples  $t_1$  and  $t_2$  satisfies the similarity condition of an MD,  $T(t_1, A, t_2, B)$  for every pair  $(A, B)$  of matched attributes in the MD.  $\square$

*Corollary 1.* Let  $D$  be an instance and  $M$  a set of non-interacting MDs. Let  $T$  be the transitive closure of the set of tuple-attribute closures of the MDs in  $M$ . Then the set of MRIs is obtained by setting, for each equivalence class  $E$  of  $T$ , the value of each attribute in  $E$  to a value that occurs in  $E$  at least as frequently as any other value in  $E$ .  $\square$

**Proof of Theorem 4:** We express the query in the form

$$\mathcal{Q}(\bar{y}) = \exists \bar{z} Q_1(\bar{z}, \bar{y}) \quad (11)$$

Let  $x_{ij}$  denote the variable of  $\bar{z}$  or  $\bar{y}$  which holds the value of the  $j^{\text{th}}$  attribute in the  $i^{\text{th}}$  conjunct  $R_i$  in  $Q_1$ . Denote this attribute by  $A_{ij}$ . Note that, since variables and conjuncts can be repeated, it can happen that  $x_{ij}$  is the same variable as  $x_{kl}$  for  $(i, j) \neq (k, l)$ , that  $A_{ij}$  is the same attribute as  $A_{kl}$  for  $(i, j) \neq (k, l)$ , or that  $R_i$  is the same as  $R_j$  for  $i \neq j$ . Let  $B$  and  $F$  denote the set of bound and free variables in  $Q_1$ , respectively. Let  $C$  and  $U$  denote the variables in  $Q_1$  holding the values of changeable and unchangeable attributes, respectively. Let  $\mathcal{Q}'(\bar{y})$  denote the rewritten query returned by algorithm *Rewrite*, which we express as

$$\mathcal{Q}'(\bar{y}) = \exists \bar{z} Q'_1(\bar{z}, \bar{y})$$

We show that, for any constant vector  $\bar{a}$ ,  $\mathcal{Q}'(\bar{a})$  is true for an instance  $D$  iff  $\mathcal{Q}(\bar{a})$  is true for all MRIs of  $D$ .



Suppose that  $\mathcal{Q}'(\bar{a})$  is true for an instance  $D$ . Then there exists a  $\bar{b}$  such that  $\mathcal{Q}'_1(\bar{a}, \bar{b})$ . We will refer to this assignment of constants to variables as  $A_{\mathcal{Q}'}$ . From the form of  $\mathcal{Q}'$ , it is apparent that, for any fixed  $i$ , there is a tuple  $t_1 = \bar{c}_i \equiv (c_{i1}, c_{i2}, \dots, c_{ip})$  such that  $R_i(\bar{c}_i)$  is true in  $D$  with the following properties.

1. For all  $x_{ij}$  except those in  $F \cap C$ ,  $c_{ij}$  is the value assigned to  $x_{ij}$  by  $A_{\mathcal{Q}'}$ .
2. For all  $x_{ij} \in F \cap C$ , there is a tuple  $t_2$  with attribute  $B$  such that  $T(t_1, A_{ij}, t_2, B)$ , where  $T$  is the transitive closure of the tuple-attribute closures of the MDs in  $M$ , such that the value of  $t_2[B]$  is the value assigned to  $x_{ij}$  by  $A_{\mathcal{Q}'}$ . Moreover, this value occurs more frequently than that of any other tuple/attribute pair in the same equivalence class of  $T$ .

For any given MRI  $D'$ , consider the tuple  $t'_1$  in  $D'$  with the same identifier as  $t_1$ . Clearly, this tuple will have the same values as  $t_1$  for all unchangeable attributes, which by 1., are the values assigned to the variables  $x_{ij} \in U$ . Also, by 2. and Corollary 1, for any  $j$  such that  $x_{ij} \in F \cap C$  is free, the value of the  $j^{th}$  attribute of  $t'_1$  is that assigned to  $x_{ij}$  by  $A_{\mathcal{Q}'}$ .

Thus, for each MRI  $D'$ , there exists an assignment  $A_{\mathcal{Q}}$  of constants to the  $x_{ij}$  that makes  $\mathcal{Q}$  true, and this assignment agrees with  $A_{\mathcal{Q}'}$  on all  $x_{ij} \notin B \cap C$ . This assignment is consistent in the sense that, if  $x_{ij}$  and  $x_{kl}$  are the same variable, they are assigned the same value. Indeed, for  $x_{ij} \notin B \cap C$ , consistency follows from the consistency of  $A_{\mathcal{Q}'}$ , and for  $x_{ij} \in B \cap C$ , it follows from the fact that the variable represented by  $x_{ij}$  occurs only once in  $\mathcal{Q}$ , by assumption. Therefore,  $\mathcal{Q}(\bar{a})$  is true for all MRIs  $D'$ , and  $\bar{a}$  is a resolved answer.

Conversely, suppose that a tuple  $\bar{a}$  is a resolved answer. Then, for any given MRI  $D'$  there is a satisfying assignment  $A_{\mathcal{Q}}$  to the variables in  $\mathcal{Q}$  such that  $\bar{z}$  as defined by (11) is assigned the value  $\bar{a}$ . We write  $\mathcal{Q}'$  in the form

$$\mathcal{Q}'(\bar{y}) \leftarrow \exists \bar{z} \wedge_{1 \leq i \leq n} Q_i(\bar{v}_i) \quad (12)$$

with  $Q_i$  the rewritten form of the  $i^{th}$  conjunct of  $\mathcal{Q}$ . For any fixed  $i$ , let  $t' = (c'_{i1}, c'_{i2}, \dots, c'_{ip})$  be a tuple in  $D'$  such that  $c'_{ij}$  is the constant assigned to  $x_{ij}$  by  $A_{\mathcal{Q}}$ .

We construct a satisfying assignment  $A_{\mathcal{Q}'}$  to the free and existentially quantified variables of  $\mathcal{Q}'$  as follows. Consider the conjunct  $Q_i$  of  $\mathcal{Q}'$  as given on line 16 of *Rewrite*. Assign to  $\bar{v}'_i$  the tuple  $t$  in  $D$  with the same identifier as  $t'$ . This fixes the values of all the variables except those  $x_{ij} \in F \cap C$ , which are set to  $c'_{ij}$ . It follows from Corollary 1 that  $A_{\mathcal{Q}'}$  satisfies  $\mathcal{Q}'$ . Since  $A_{\mathcal{Q}}$  and  $A_{\mathcal{Q}'}$  match on all variables that are not local to a single  $Q_i$ ,  $A_{\mathcal{Q}'}$  is consistent. Therefore,  $\bar{a}$  is an answer for  $\mathcal{Q}'$  on  $D$ .  $\square$

**Proof of Theorem 5:** Hardness follows from the fact that, for the instance  $D$  resulting from the reduction in the proof of Theorem 3.3 in [12], the set of all repairs of  $D$  with respect to the given key constraint is the same as the set of MRIs with respect to (5). The key point is that attribute modification in this case generates duplicates which are subsequently eliminated from the instance, producing the same result as tuple deletion. Containment follows from Theorem 3.  $\square$

**Proof of Proposition 2:** Take  $\bar{A} = (A_1, \dots, A_m)$  and  $\bar{B} = (B_1, \dots, B_n)$ . For any tuple of constants  $\bar{k}$ , define  $R^{\bar{k}} \equiv$

$\sigma_{\bar{A}=\bar{k}} R$ . Let  $B_i^{\bar{k}}$  denote the single attribute relation with attribute  $B_i$  whose tuples are the most frequently occurring values in  $\pi_{B_i} R^{\bar{k}}$ . That is,  $a \in B_i^{\bar{k}}$  iff  $a \in \pi_{B_i} R^{\bar{k}}$  and there is no  $b \in \pi_{B_i} R^{\bar{k}}$  such that  $b$  occurs as the value of the  $B_i$  attribute in more tuples of  $R^{\bar{k}}$  than  $a$  does. Note that  $B_i^{\bar{k}}$  can be written as an expression involving  $R$  which is first order with a *Count* operator. The reduction produces  $(R', t)$  from  $(R, t)$ , where

$$R' \equiv \bigcup_{\bar{k}} \left[ \pi_{\bar{A}} R^{\bar{k}} \times B_1^{\bar{k}} \times \dots \times B_n^{\bar{k}} \right] \quad (13)$$

The repairs of  $R'$  are obtained by keeping, for each set of tuples with the same key value, a single tuple with that key value and discarding all others. By Corollary A.1, in a MRI of  $D$ , the group  $G_{\bar{k}}$  of tuples such that  $\bar{A} = \bar{k}$  for some constant  $\bar{k}$  has a common value for  $\bar{B}$  also, and the set of possible values for  $\bar{B}$  is the same as that of the tuple with key  $\bar{k}$  in a repair of  $D$ . Since duplicates are eliminated from the MRIs, the set of MRIs of  $D$  is exactly the set of repairs of  $R'$ .  $\square$

**Proof of Theorem 6:**  $\mathcal{Q}''$  is obtained by composing  $\mathcal{Q}'$  with the transformation  $R \rightarrow R'$ , which is a first-order query with aggregation.  $\square$